A UNITED STATES
DEPARTMENT OF
COMMERCE
PUBLICATION

National Bureau of Standards
Library, E-01 Admin. Bldg.

FEB 6 1970

NBS TECHNICAL NOTE 492

copy 2

Component Combination and Frame-Embedding in Chinese Character Grammars

U.S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress March 3, 1901. Today, in addition to serving as the Nation's central measurement laboratory, the Bureau is a principal focal point in the Federal Government for assuring maximum application of the physical and engineering sciences to the advancement of technology in industry and commerce. To this end the Bureau conducts research and provides central national services in four broad program areas. These are: (1) basic measurements and standards, (2) materials measurements and standards, (3) technological measurements and standards, and (4) transfer of technology.

The Bureau comprises the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, the Center for Radiation Research, the Center for Computer Sciences and Technology, and the Office for Information Programs.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement; coordinates that system with measurement systems of other nations; and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of an Office of Measurement Services and the following technical divisions:

Applied Mathematics—Electricity—Metrology—Mechanics—Heat—Atomic and Molecular Physics—Radio Physics[2]—Radio Engineering[2]—Time and Frequency[2]—Astrophysics[2]—Cryogenics.[2]

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to improved methods of measurement standards, and data on the properties of well-characterized materials needed by industry, commerce, educational institutions, and Government; develops, produces, and distributes standard reference materials; relates the physical and chemical properties of materials to their behavior and their interaction with their environments; and provides advisory and research services to other Government agencies. The Institute consists of an Office of Standard Reference Materials and the following divisions:

Analytical Chemistry—Polymers—Metallurgy—Inorganic Materials—Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides technical services to promote the use of available technology and to facilitate technological innovation in industry and Government; cooperates with public and private organizations in the development of technological standards, and test methodologies; and provides advisory and research services for Federal, state, and local government agencies. The Institute consists of the following technical divisions and offices:

Engineering Standards—Weights and Measures — Invention and Innovation — Vehicle Systems Research—Product Evaluation—Building Research—Instrument Shops—Measurement Engineering—Electronic Technology—Technical Analysis.

THE CENTER FOR RADIATION RESEARCH engages in research, measurement, and application of radiation to the solution of Bureau mission problems and the problems of other agencies and institutions. The Center consists of the following divisions:

Reactor Radiation—Linac Radiation—Nuclear Radiation—Applied Radiation.

THE CENTER FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides technical services designed to aid Government agencies in the selection, acquisition, and effective use of automatic data processing equipment; and serves as the principal focus for the development of Federal standards for automatic data processing equipment, techniques, and computer languages. The Center consists of the following offices and divisions:

Information Processing Standards—Computer Information — Computer Services — Systems Development—Information Processing Technology.

THE OFFICE FOR INFORMATION PROGRAMS promotes optimum dissemination and accessibility of scientific information generated within NBS and other agencies of the Federal government; promotes the development of the National Standard Reference Data System and a system of information analysis centers dealing with the broader aspects of the National Measurement System, and provides appropriate services to ensure that the NBS staff has optimum accessibility to the scientific information of the world. The Office consists of the following organizational units:

Office of Standard Reference Data—Clearinghouse for Federal Scientific and Technical Information[3]—Office of Technical Information and Publications—Library—Office of Public Information—Office of International Relations.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.

[2] Located at Boulder, Colorado 80302.

[3] Located at 5285 Port Royal Road, Springfield, Virginia 22151.

UNITED STATES DEPARTMENT OF COMMERCE
Maurice H. Stans, Secretary

NATIONAL BUREAU OF STANDARDS • Lewis M. Branscomb, Director

NBS TECHNICAL NOTE 492

ISSUED FEBRUARY 1970

Nat. Bur. Stand. (U.S.), Tech. Note 492, 36 pages (Feb. 1970)
CODEN: NBTNA

Component Combination and Frame-Embedding in Chinese Character Grammars

Kirk Rankin and James L. Tan

Systems Development Division
Center for Computer Sciences and Technology
National Bureau of Standards
Washington, D.C. 20234

NBS Technical Notes are designed to supplement the Bureau's regular publications program. They provide a means for making available scientific data that are of transient or limited interest. Technical Notes may be listed or referred to in the open literature.

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 20402
(Order by SD Catalog No. C13.46:492), Price 40 cents

Contents

Page

1. Introduction 1

2. The Grammar 7

    2.1 The Syntax 8

    2.2 The Lexicon 12

    2.3 Variation 28

3. Concluding Remarks 32

4. References 34

Component Combination and Frame-embedding in Chinese Character Grammars

Kirk Rankin and James L. Tan

Chinese characters can be almost completely described from the point of view that each character is composed of a number of components and that each component is composed of a number of strokes. This note offers a grammatical treatment of component combination. It is concerned with the three most productive processes of component arrangement. This grammar differs from previous grammars in that the constraints on recursion have been minimized and great lexical economy has been achieved.

Key words: Chinese characters; grammar; generative grammar; component combination; linguistics; frame-embedding; blocking.

1. Introduction

This note is one of a series which is generally concerned with the linguistic study of the formation of Chinese characters.[1] It offers a particular treatment of one aspect of the general study. In this preface we wish to establish a frame of reference for placing the treatment offered here within the general study. To that end we will initiate a discussion of the formation of Chinese characters and will successively focus our attention on less general aspects of it until the treatment in this note is reasonably well characterized.

Concerning the formation of Chinese characters the following comments can be made. It has been observed that the internal structure of Chinese characters bears certain resemblances to the internal structure of natural language sentences.[2] For example, sentences can be viewed as being composed of word sequences and ultimately of words or word-parts, and words and word-parts can be viewed as being sequences of sound units. Analogously, Chinese characters can be almost completely described as being composed of a number of components, components in turn being composed of a number of strokes. Thus 後 , a character, is composed of three components (彳 , 幺 , and 夂). In turn, 彳 is composed of the strokes 丿 , 丿 , and 丨 ; 幺 is composed of the strokes く , ㄥ , and 丶 ; and 夂 is composed of the strokes 丿 , フ , and ㇏ .

---

[1] Burkart (1), Rankin (3), Rankin et al.(4), Rankin et al.(5) (see references) are the other publications in the series.

This note is restricted to the area of component combination, the manner in which characters are composed of components. Examples of characters analyzed in terms of components are:

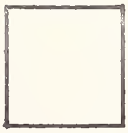
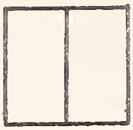
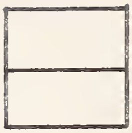
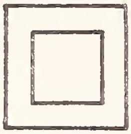
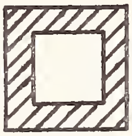
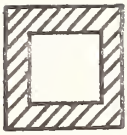
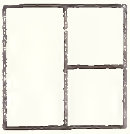
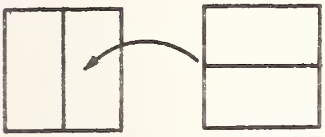
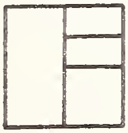
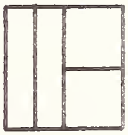
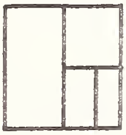
Character	Component(s)
木	木
相	木, 目
亢	亠, 几
困	口, 木
近	辶, 斤
庫	广, 車

---

[2] See Rankin (3), especially Chapter VIII, for a detailed discussion of this point.

Character	Component(s)
可	丁, 口
後	彳, 幺, 夂
清	氵, 龶, 月
路	口, 止, 夂, 口
絍	幺, …, 亻, 壬
高	亠, 口, 冂, 口

The concept of component combination has been informally introduced. We must now face the problem of how to account for it in a linguistic analysis of Chinese characters. One way to account for it (and this is by no means the only way) is to construct a grammar to generate output objects which correspond to Chinese characters using components as the terminal symbols in the grammar. This is the approach followed in all the previous papers in this series except Burkart(1). Given that we wish to pursue the generative grammar approach the question then arises of what mechanism will enable us to characterize the spatial combination of components into characters. One way to represent it is by means of the process of frame-embedding. That process will now be introduced.

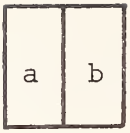
First,  is a frame, and represents all single-component characters, such as 木. Second, (1) , (2) , and (3)  are frames. (1) represents those two-component characters, like 相, in which the two components, (木 and 目 in this case) are horizontally arranged. (2) represents those two-component characters, like 亢, in which the two components (亠 and 几 in this case) are vertically arranged. (3) represents those two-component characters, like 困, in which the two components (口 and 木 in this case) are arranged in a surrounding/surrounded fashion. 困 is an instance of total surrounding; (3) also represents instances of partial surrounding, like 近, 庫, and 可. Frames for characters of component complexity greater than two are derived by means of the embedding of (1), (2), (3) in any subframe of (1), (2), (3) except the shaded subframe in (3): . The process of frame-embedding is recursive. That is, once a frame has been embedded in a subframe, any subframe of the derived frame (except ) may be embedded in, and so on without limit. Thus, for example, once  is derived from  any of its subframes can be embedded in, yielding, for example, , , or , etc.

Thus there are arbitrarily "large" frames in terms of the number of embeddings. This does not imply that there are arbitrarily large characters in terms of component-number. Our position is that there is an upper bound on the number of components per character[3], hence on frame size, but that (1) the precise location of the upper bound is not known and (2) the grammar would be less elegant if the upper bound were built in via, e.g., a loop counter.

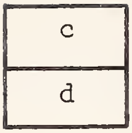
There are presumably many kinds of restrictions we can impose on the above stated process. We could thus build a theory of frame-embedding which would be somewhat analogous to the theory of string languages, but that is not our purpose here. However, one restriction does have interesting results. Let us define each of the following pair of subframes as "partner" subframes:

---

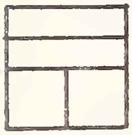
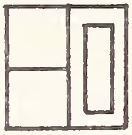
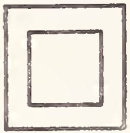
[3] In a sample of 400 characters from Mathews' (2), we found that the upper bound was 9.

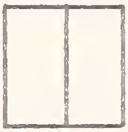
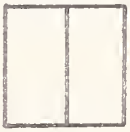
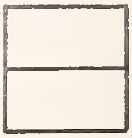


pair 1 (a,b)



pair 2 (c,d)

Now if at embedding time the partner of the "embedded in" subframe is blocked from future embedding, we can no longer derive such frames as  and , or generally any frame where there is embedding in both partner subframes of a pair. This restriction is worth mentioning because it is adhered to in the grammars in Rankin (3) and Rankin et. al.(4). Note that the border of  is always blocked, both in past work and in this presentation.

The blocking restriction has an important implication for the size of the lexicon. There are many characters which are complex in both partner subframes of a pair. Examples are: (1) 韶 , which is complex in both partner subframes of  ; (2) 髞 , which is complex in both partner subframes of  , and which further contains 高 , which is complex in both partner subframes of  . Such multiple complexity is quite common, and for a blocked grammar to generate such characters, it must list at least one of the complex structures in its lexicon. In the case of 韶 above, either 音 or 召 must appear in the lexicon, and obviously both are composed of components already in the lexicon: 音 = 亠, 䒑, 日 ; and 召 = 刀 , 口 . Consequently, lexicons in blocked grammars necessarily contain more entries than those in (partially) unblocked grammars.

So we see that there are two versions of frame embedding which have been used in grammars for Chinese characters: the blocked version (used previously) and the partially umblocked version (used in the current treatment). Given any version of frame-embedding there are many grammars based upon it which all derive their common syntactic power (i.e., their ability to impose certain kinds of constituent structure on generated characters) from the process of frame-embedding. They would differ from each other primarily in what components are listed in the lexicon and in how the components are classified. We now present one grammar based on the partially unblocked version of frame-embedding.

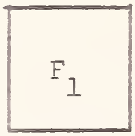
2. The Grammar

In this section we present a grammar capable of generating all the entries in Mathews' (2) minus two small classes of entries which are deliberately excluded. The first class of excluded entries are those traditional radicals which are not characters; that is, they are always bound, such as 氵 and 刂 . The second class consists of uncommon forms of characters which are more commonly represented by variant forms which are generated by the grammar. Here we refer to such uncommon forms as 爲 and 者, which are usually replaced by 為 and 者 in common practice.

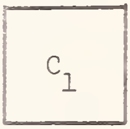
The grammar comes in two parts: a syntax and a lexicon. The two parts are linked by the use of component class marks. That is, the syntax generates all possible frames, with subframes filled by component class marks. The lexicon lists all components and marks them as members of certain of these classes. The lexicon further indicates certain component variation information. The outputs of lexical selection are frames filled with components -- our representation of Chinese characters. It is understood that components in distinct subframes do not touch each other.

2.1 The Syntax

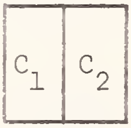
I. CHAR →



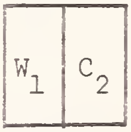
II. CHAR →



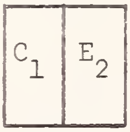
III. C →



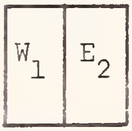
IV. C →



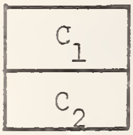
V. C →



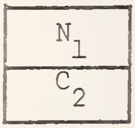
VI. C →



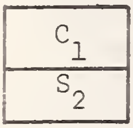
VII. C →



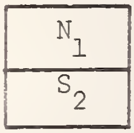
VIII. C →



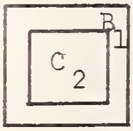
IX. C →



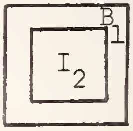
X. C →



XI. C →



XII. C →

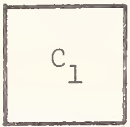
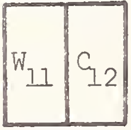
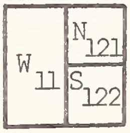


The syntax consists of twelve replacement rules (I, II, ..., XII) of the form p → q, where p is a non-terminal symbol (CHAR or C), and q is either a frame filled with one subscripted symbol or a frame filled with two subscripted symbols, terminal, non-terminal or mixed. The subscripts can be ignored for now. The terminal symbols are W=WEST, E=EAST, N=NORTH, S=SOUTH, B=BORDER, I=INTERIOR, and F=FREE. These terminal symbols are terminal with respect to the syntax. With respect to the total grammar, they are "sub-terminal" symbols, i.e., those symbols which are replaced by lexical items. The initial symbol of the syntax is CHAR (for CHARACTER), and every generation begins, by convention, with that symbol.

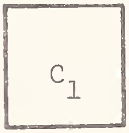
The generation of a terminal frame (i.e., a frame whose subframes are filled with terminal symbols) is as follows. A sequence of stages is created such that stage one is always the initial symbol CHAR, and stage two is either $\boxed{C_1}$ or $\boxed{F_1}$. In general, stage i + 1 is developed from stage i by the replacement of an occurence of C in stage i by any q of any rule in which p = C. When there are more than one occurrences of C in a stage, the question of which occurrence is replaced is determined by the subscript convention given below. The process continues until stage i contains no occurrences of C, or equivalently, until every subframe is filled with a terminal syntactic symbol.

Note that there is a subscript on each symbol occupying a subframe in the q portion of every rule. With each replacement of C, the subscripts on C are carried along to the symbols in q, and the subscript on the symbols in q is concatenated to the right of the subscript carried along from C. At any stage there may be several occurrences of C with subscripts. The C with the lowest valued subscript in lexicographic ordering is the next symbol to be replaced. For instance, 111 is lower valued than 12, etc. To determine the lowest-valued of two subscripts, the following test is made. At some position in a left to right scan of digit positions, there will be different digits, a 1 and a 2. The subscript containing the 1 at this position is the lowest valued. Example of generations of terminal frames are now given.

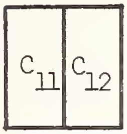
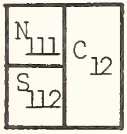
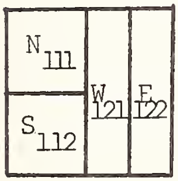
Example 1 (for the terminal frame which corresponds to 清)

	Stages	Authority
1.	CHAR	Initial Symbol Convention
2.		Rule II
3.		Rule IV
4.		Rule X

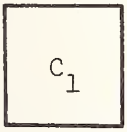
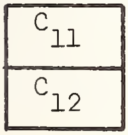
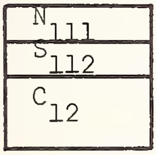
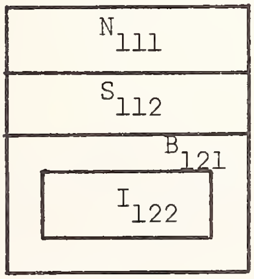
Example 2 (for the terminal frame which corresponds to 紝)

	Stages	Authority
1.	CHAR	Initial Symbol Convention
2.		Rule II

Example 2 (continued)

	Stages	Authority
3.		Rule III
4.		Rule X
5.		Rule VI

Example 3 (for the terminal frame which corresponds to 高)

	Stages	Authority
1.	CHAR	Initial Symbol Convention
		Rule II
		Rule VII
		Rule X
		Rule XIII

2.2 The Lexicon

The lexicon is a table with components heading the rows and component class marks heading the columns. An x at a row/column intersection point indicates that the component heading that row is a member of the class heading that column, and thus may replace that class mark in any terminal frame. Note that some occurrences of x are superscripted; these numbers indicate that a variation process is to be effected as the component heading the row replaces the class mark heading the column. The variation processes are stated on pages 28 to 31 immediately following the lexicon.

Lexical rules can be viewed as being of the same form as syntactic rules, namely $p \rightarrow q$, with the restriction that p is a component class mark and q is a component. The lexicon is thus an abbreviation of a very large number of such rules -- the number, in fact, of x's in the lexicon.

Given a terminal frame, the application of lexical rules is as follows. The subscripts on Σ , the component class marks, in the terminal frame are lexicographically ordered. The Σ with the lowest-valued subscript is considered first. The column headed by Σ is scanned and any component heading a row which has an x in Σ replaces Σ in the terminal frame. There is now a unique lowest valued subscript. The Σ associated with that subscript is now considered and an appropriate lexical replacement is effected. The process continues until every Σ is replaced by a component.[4]

---

[4] Note that the use of subscripts results in an infinite number of nonterminal symbols. We are using these only as a notation device to reflect the history of the derivation. A detailed discussion of this would be appropriate in a paper concerned with theoretical issues.

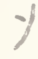
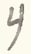
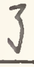
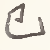
The lexicon is now displayed. The components are numbered according to the format S.T.N. where S indicates the total number of strokes in the component, T the type of the last stroke in the component, and N the number of the component in the sub-list determined by S and T. The values of T are 1, 2, ..., 8 where 1 = horizontal, 2 = vertical, 3 = dot, 4 = \\-like, 5 = /-like, 6 = /-like, 7 = hooked, and 8 = multi-directional.[5]

---

[5] 5 and 6 differ in direction. 5 is northeast to southwest, and 6 is southwest to northeast.

NUM	CHAR	W	E	N	S	B	I	F
1.1.1	一			X[5]	X		X	X
1.2.1	丨		X					
1.7.1	亅							X
1.7.2	乛			X				
1.8.1	乚		X				X	X
1.8.2	乙				X	X[5]	X	X
1.8.3	𠃊				X	X		
1.8.4	㇆		X			X		
1.8.5	𡿨	X	X					
2.1.1	⊢			X				
2.1.2	二		X	X	X	X[5]	X	X
2.1.3	ユ			X				
2.1.4	丄			X				
2.1.5	コ			X				
2.1.6	ニ			X				
2.1.7	ト			X				
2.2.1	丩	X	X				X	X
2.2.2	卩		X		X			X
2.2.3	十	X	X	X	X	X[5]		X

NUM	CHAR	W	E	N	S	B	I	F
2.2.4	凵					X		
2.2.5	亻	X						
2.2.6	丬	X	X					X
2.2.7	刂	X			X			
2.2.8	丅			X	X			
2.3.1	厶	X	X	X	X		X	X
2.3.2	卜		X		X		X	X
2.3.3	マ			X				
2.3.4	丷			X	X		X	
2.3.5	冫				X			
2.4.1	又	X	X	X	X		X	X
2.4.2	乂	X		X[5]	X[5]		X	X
2.4.3	人	X	X	X[5]	X		X	X
2.4.4	八		X	X	X			X
2.5.1	刀	X	X	X	X			X
2.5.2	力	X	X	X	X		X	X
2.5.3	入						X[5]	X
2.5.4	丿					X		
2.5.5	ソ			X				
2.5.6	厂					X		

NUM	CHAR	W	E	N	S	B	I	F	NUM	CHAR	W	E	N	S	B	I	F
2.5.7	九	X	X		X	X[1]	X	X	2.8.7	匚					X		
2.5.8		X					X		2.8.8	丂	X	X		X		X	X
2.5.9	厂					X			2.89				X				
2.5.10							X		2.8.10					X			X
2.5.11				X		X[5]			2.8.11		X	X	X	X		X	X
2.5.12		X							2.8.12	七			X	X		X	X
2.5.13				X		X[5]	X		2.8.13		X	X	X	X		X	X
2.6.1	刁		X		X			X	2.8.14	乃		X	X	X	X[5]		X
2.6.2		X			X				2.8.15		X						
2.7.1	丁	X	X		X	X[5]	X	X	2.8.16					X			
2.7.2				X	X				2.8.17	与					X		
2.7.3	刂		X														
2.7.4					X			X	3.1.1	士	X	X	X				X
2.7.5	了				X				3.1.2	土	X[2]	X	X	X		X	X
2.8.1		X	X	X	X		X	X	3.1.3	工	X[2]	X	X	X		X	X
2.8.2	几		X	X	X	X[1]	X	X	3.1.4	子	X[5]	X	X	X		X	X
2.8.3	勹			X		X		X	3.1.5		X						
2.8.4								X	3.1.6	口	X	X	X	X	X[5]	X	X
2.8.5	冂				X	X		X	3.1.7	上			X				X
2.8.6	儿				X		X	X	3.1.8				X	X			

NUM	CHAR	W	E	N	S	B	I	F	NUM	CHAR	W	E	N	S	B	I	F
3.1.9	三				X			X	3.3.1	弋		X	X		X[5]		X
3.1.10	乇						X	X	3.3.2	叉		X					X
3.1.11	女	X	X	X	X		X	X	3.3.3	幺	X	X	X	X			X
3.1.12	⺕			X[5]	X				3.3.4	刃	X	X	X			X	X
3.1.13	五			X					3.3.5	丸		X	X			X	X
3.1.14	𠂉			X					3.3.6	凢		X	X				X
3.1.15	⺊				X				3.3.7	寸		X		X		X	X
3.2.1	巾	X	X		X		X	X	3.3.8	凡		X	X	X			X
3.2.2	千	X[5]	X	X	X	X[5]		X	3.3.9	云			X	X			
3.2.3	干	X	X	X	X[5]	X[5]	X	X	3.3.10	勺		X		X			X
3.2.4	阝	X	X					X	3.3.11	夕	X	X	X	X		X	X
3.2.5	山	X	X	X	X		X	X	3.3.12	厶	X						
3.2.6	扌				X		X[5]	X	3.3.13	下		X	X	X			X
3.2.7	卄			X	X[5]		X	X	3.3.14	乍	X	X	X				X
3.2.8	丫	X						X	3.3.15	小		X	X	X		X	X
3.2.9	亻	X							3.3.16	犬					X		
3.2.10	彳	X							3.3.17	么	X	X		X		X	X
3.2.11	才					X			3.3.18	忄	X						
3.2.12	卜		X						3.3.19	文				X			
3.2.13	午	X							3.3.20	⺍	X						

NUM	CHAR	W	E	N	S	B	I	F
3.3.21	儿				X			
3.3.22	巛				X			
3.3.23	冈				X			
3.4.1	大	X	X	X[5]	X		X	X
3.4.2	丈		X					X
3.4.3	欠		X		X		X	X
3.4.4	廴					X		
3.4.5	丞							X
3.4.6	以				X			
3.4.7	夂			X	X	X[4]		X
3.4.8	氏		X					
3.5.1	才		X				X	X
3.5.2	尸	X				X		X
3.5.3	彡	X	X	X	X		X	X
3.5.4	乡	X						
3.5.5	勿		X					X
3.5.6	广					X		
3.5.7	屮	X	X	X[5]	X		X	X
3.5.8	歹	X						
3.5.9	万				X			
3.5.10	少				X			
3.5.11	⺌			X				
3.5.12	为						X	
3.6.1	扌	X						
3.6.2	氵	X						
3.6.3	子	X						X
3.7.1	于		X	X	X		X	X
3.7.2	宀			X				
3.7.3	宁		X					
3.8.1	兀		X		X	X[1]		X
3.8.2	尢			X		X[1]		X
3.8.3	已		X					X
3.8.4	巳	X	X	X			X	X
3.8.5	己	X	X	X				X
3.8.6	乇	X	X	X	X		X	X
3.8.7	弓	X	X		X	X[5]		X
3.8.8	门					X		
3.8.9	凡		X					X
3.8.10	亏		X		X			X
3.8.11	也		X		X		X	X

NUM	CHAR	W	E	N	S	B	I	F
3.8.12	丂					X		
4.1.1	曰							X
4.1.2	[illegible]			X	X			X
4.1.3	冃	X						X
4.1.4	丑		X				X	X
4.1.5	月	X	X	X	X^{5}		X	X
4.1.6	止	X^{2}	X	X	X		X	X
4.1.7	[illegible]			X				
4.1.8	王	X^{2}	X	X	X		X	X
4.1.9	丗			X				
4.1.10	五		X	X				X
4.1.11	生			X				X
4.1.12	壬		X		X		X	X
4.1.13	廿			X				X
4.1.14	互		X					X
4.1.15	丹	X	X				X	X
4.1.16	毋			X			X	X
4.1.17	白				X	X		
4.1.18	日^{5}	X	X	X	X		X	X
4.1.19	主	X	X	X	X			X
4.1.20	毋							X
4.1.21	四			X				
4.2.1	牛	X	X	X	X		X	X
4.2.2	开	X	X		X		X	X
4.2.3	冊				X			
4.2.4	井		X		X			X
4.2.5	中		X	X			X	X
4.2.6	斤	X	X		X		X	X
4.2.7	弔	X	X					X
4.2.8	午		X^{5}		X			X
4.2.9	申			X				X
4.2.10	升			X	X			X
4.2.11	州							X
4.2.12	币		X		X		X	X
4.2.13	丯	X	X	X^{5}	X^{5}			X
4.2.14	斗		X		X		X	X
4.2.15	冃				X	X^{5}	X^{5}	
4.2.16	手				X			
4.2.17	丰	X	X					

NUM	CHAR	W	E	N	S	B	I	F	NUM	CHAR	W	E	N	S	B	I	F
4.3.1	心	X	X	X	X		X	X	4.4.4	夭		X	X	X			X
4.3.2	内		X		X	X		X	4.4.5	天		X	X	X		X	X
4.3.3	太		X	X	X			X	4.4.6	𧘇				X			
4.3.4	不		X	X	X		X	X	4.4.7	水		X	X	X		X	X
4.3.5	歹	X			X		X	X	4.4.8	𧘇				X[5]		X	X
4.3.6	戈	X	X	X	X	X[5]	X	X	4.4.9	之				X			X
4.3.7	犬	X	X	X	X		X	X	4.4.10	辶					X		
4.3.8	叉			X				X	4.4.11	不							X
4.3.9	于				X			X	4.4.12	尺				X	X[4]	X	X
4.3.10	尤		X	X	X	X[1]	X	X	4.4.13	木	X[3]	X	X	X[5]		X	X
4.3.11	夕	X		X				X	4.4.14	爪		X		X	X[4]		X
4.3.12	小				X			X	4.4.15	夊		X					X
4.3.13	灬				X		X		4.4.16	火	X	X	X	X		X	X
4.3.14	小				X				4.4.17	止				X			
4.3.15	幺			X					4.5.1	尹		X		X	X		X
4.3.16	不				X				4.5.2	户	X	X			X		X
4.3.17	弋					X			4.5.3	东		X					X
4.4.1	夫	X	X		X		X	X	4.5.4	片	X						X
4.4.2	夬	X	X	X				X	4.5.5	幺			X				
4.4.3	及		X		X			X	4.5.6	牙	X	X		X		X	X

NUM	CHAR	W	E	N	S	B	I	F
4.5.7	尸				X	X		X
4.5.8	歹					X[5]	X	X
4.5.9	夬					X		
4.5.10	少	X	X	X		X[5]		X
4.5.11	丰	X						
4.5.12	手				X			
4.5.13	手	X				X		
4.6.1	牛	X						
4.7.1	氏		X		X			
4.7.2	手		X		X			
4.7.3	聿				X			
4.8.1	元	X	X	X	X	X[1]	X	X
4.8.2	旡	X	X	X				X
4.8.3	巴		X		X		X	X
4.8.4	屯	X	X		X		X	X
4.8.5	冘	X	X	X				X
4.8.6	毛	X	X	X	X	X[1]	X	X
4.8.7	公					X		
4.8.8	无							X
4.8.9	巳						X	X

NUM	CHAR	W	E	N	S	B	I	F
4.8.10	门				X			
4.8.11	方	X	X		X		X	X
4.8.12	丐		X					X
4.8.13	丐		X					X
4.8.14	片	X						X
5.1.1	虫			X				X
5.1.2	丘	X	X	X				X
5.1.3	圭			X	X			X
5.1.4	目	X	X	X	X		X	X
5.1.5	且	X	X		X		X	X
5.1.6	巨		X		X			X
5.1.7	吕		X		X			X
5.1.8	互		X					X
5.1.9	册	X	X					X
5.1.10	廿	X	X	X	X		X	X
5.1.11	正	X[2]	X	X	X		X	X
5.1.12	企						X	X
5.1.13	聿			X				X
5.1.14	业			X	X		X	X

NUM	CHAR	W	E	N	S	B	I	F
5.1.15	生	X[2]	X	X	X		X	X
5.1.16	非			X	X			
5.1.17	凹			X				X
5.1.18	凸							X
5.1.19	本		X		X			X
5.1.20	疋	X[2]						X
5.1.21	冉		X		X			X
5.1.22	四		X	X	X			X
5.1.23	皿				X			X
5.1.24	白	X	X	X	X		X	X
5.1.25	由		X	X	X		X	X
5.1.26	母		X	X	X			X
5.1.27	田	X	X	X	X		X	X
5.1.28	巨			X				
5.1.29	[illegible]	X						
5.2.1	申		X					X
5.2.2	弗		X	X	X		X	X
5.2.3	半	X	X					X
5.2.4	甲	X	X	X[5]			X	X
5.2.5	出		X	X	X		X	X

NUM	CHAR	W	E	N	S	B	I	F
5.2.6	聿			X				
5.2.7	平		X		X		X	X
5.2.8	[illegible]				X			
5.2.9	手		X					
5.2.10	申			X				
5.2.11	卯		X		X	X[5]		X
5.3.1	朮		X				X	X
5.3.2	民	X						X
5.3.3	氏	X	X				X	X
5.3.4	水				X			X
5.3.5	玉				X			X
5.3.6	戊				X	X[5]		X
5.3.7	瓦		X		X			X
5.3.8	斥		X					X
5.3.9	丙	X	X	X	X		X	X
5.3.10	礻	X						
5.3.11	必		X	X	X		X	X
5.3.12	去				X			
5.3.13	内				X			
5.3.14	太			X				

NUM	CHAR	W	E	N	S	B	I	F
5.4.1	皮	X	X		X		X	X
5.4.2	未	X	X				X	X
5.4.3	末		X		X			X
5.4.4	氶			X				X
5.4.5	瓜	X	X		X	X[4]	X	X
5.4.6	央		X	X	X			X
5.4.7	永		X	X		X[4]		X
5.4.8	失		X		X		X	X
5.4.9	水				X			
5.4.10	反			X			X	X
5.4.11	吏			X			X	X
5.4.12	禾				X			X
5.4.13	疋				X			X
5.4.14	奴			X			X	X
5.4.15	夫			X				X
5.4.16	乏		X		X			X
5.4.17	史		X					X
5.4.18	乐	X[3]	X	X	X		X	X
5.4.19	飞			X				
5.5.1	弗				X			X
5.5.2	户					X		
5.6.1	乒	X						X
5.6.2	疒					X		
5.7.1	民	X	X	X				X
5.7.2	乎		X				X	X
5.8.1	电				X			X
5.8.2	包	X	X		X		X	X
5.8.3	电				X			X
5.8.4	寿	X	X					X
5.8.5	世		X	X			X	X
5.8.6	甩							X
5.8.7	为					X		
5.8.8	馬					X		
5.8.9	北	X		X		X[5]		X
6.1.1	自		X	X	X			X
6.1.2	臼		X	X	X		X	X
6.1.3	血	X[2]	X	X				X
6.1.4	聿			X				X
6.1.5	自	X	X	X	X		X	X

NUM	CHAR	W	E	N	S	B	I	F	NUM	CHAR	W	E	N	S	B	I	F
6.1.6	亚			X	X			X	6.3.2	戊		X				X	X
6.1.7	西		X	X	X		X	X	6.3.3	戉				X			X
6.1.8	再							X	6.3.4	成		X	X	X			X
6.1.9	襾		X					X	6.3.5	𢦏					X		
6.1.10	覀		X	X	X			X	6.3.6	兆	X	X	X	X		X	X
6.1.11	舟	X	X					X	6.3.7	乒							X
6.1.12	壬				X^2				6.3.8	兵							X
6.1.13	曲			X					6.3.9	虫	X^5	X	X	X		X	X
6.1.14	面			X					6.3.10	录						X	X
6.1.15	且			X					6.3.11	㠯	X			X			X
6.1.16	四				X				6.3.12	戋					X		
6.1.17	丗			X					6.3.13	※						X	
6.1.18	其			X					6.3.14	栽				X			
6.2.1	甲				X			X	6.4.1	吏		X					X
6.2.2	缶	X			X		X	X	6.4.2	東	X	X	X	X			X
6.2.3	角	X						X	6.4.3	艮		X	X			X	X
6.2.4	聿		X		X		X	X	6.4.4	朱	X	X	X	X			X
6.2.5	半				X			X	6.4.5	耒	X	X					X
6.2.6	耳	X^5	X		X	X^5	X	X	6.4.6	叀	X	X				X	X
6.3.1	衣	X							6.47	米	X	X	X	X		X	X

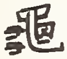
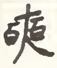
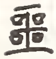
NUM	CHAR	W	E	N	S	B	I	F
6.4.8	美			X				X
6.4.9	永				X			
6.4.10	火			X				X
6.4.11	夗			X				
6.5.1	曳		X					X
6.5.2	男	X	X		X		X	X
6.5.3	甶			X				
6.7.1	且			X				
6.7.2	凹			X		X[5]		
6.8.1	身	X						X
6.8.2	臣	X	X		X		X	X
6.8.3	卌			X				
6.8.4	鸟					X		
6.8.5	冎					X		
7.1.1	酉	X	X		X			X
7.1.2	里	X[2]	X	X	X		X	X
7.1.3	缶	X						X
7.1.4	坐	X	X		X		X	X
7.1.5	巫	X	X		X			X

NUM	CHAR	W	E	N	S	B	I	F
7.1.6	齒			X				X
7.1.7	臼					X		
7.1.8	重				X			
7.1.9	西			X				
7.1.10	巫			X				
7.1.11	皀		X					
7.2.1	串			X				X
7.2.2	車	X	X	X	X		X	X
7.2.3	甲			X				
7.3.1	南	X	X		X			X
7.3.2	甫		X	X	X		X	X
7.3.3	求	X	X				X	X
7.3.4	我	X	X		X			X
7.3.5	鱼				X[5]		X	X
7.3.6	率			X				X
7.3.7	兩				X			X
7.3.8	隶				X		X	X
7.3.9	卵							X
7.4.1	夾	X	X	X	X		X	X
7.4.2	釆	X	X	X	X		X	X

NUM	CHAR	W	E	N	S	B	I	F	NUM	CHAR	W	E	N	S	B	I	F
7.4.3	足		X		X	X[4]		X	8.1.6	非		X	X	X		X	X
7.4.4	良	X	X		X		X	X	8.2.1	肅						X	X
7.4.5	更		X			X[4]		X	8.2.2	串			X				X
7.4.6	東	X[3]	X		X		X	X	8.2.3	垂				X			
7.4.7	走		X			X[4]		X	8.3.1	兩		X				X	X
7.4.8	豕	X	X		X	X[4]	X	X	8.3.2	隶		X				X	X
7.4.9	臭				X				8.3.3	重	X		X				X
7.5.1	豸	X							8.3.4	雨			X			X	X
7.5.2	身	X	X					X	8.3.5	戠					X		
7.7.1	臥				X				8.3.6	我					X		
7.8.1	見		X		X	X[1]		X	8.3.7	戕					X		
7.8.2	鳥					X			8.3.8	𢦔					X		
7.8.3	皃					X[1]	X	X	8.4.1	果	X	X	X	X			X
7.8.4	臣	X	X					X	8.4.2	承							X
									8.4.3	疌		X					X
8.1.1	亟		X					X	8.4.4	東		X					X
8.1.2	亞	X	X	X	X		X	X	8.4.5	兒	X	X				X	X
8.1.3	罡			X					8.4.6	來	X	X		X			X
8.1.4	𦥑			X		X			8.4.7	兼		X					X
8.1.5	亜	X						X	8.7.1	事	X						X

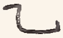
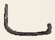
NUM	CHAR	W	E	N	S	B	I	F
8.7.2	亟			X				
8.8.1	乖							X
8.8.2	虎					X		
8.8.3	門		X			X		X
9.1.1	垂		X		X		X	X
9.1.2	重	X	X		X		X	X
9.1.3	缶	X	X					X
9.1.4	亞	X	X		X			X.
9.1.5	巠	X						X
9.2.1	崩		X					X
9.2.2	叀	X	X					X
9.3.1	飛				X			X
9.3.2	風				X	X[1]	X	X
9.3.3	禺		X	X	X		X	X
9.3.4	禹	X					X	X
9.4.1	東		X				X	X
9.4.2	是					X[4]		X
9.7.2	臼					X		

NUM	CHAR	W	E	N	S	B	I	F
10.1.1	隺	X	X					X
10.1.2	重			X				X
10.2.1	卒				X			X
10.3.1	鬼		X		X	X[1]		X
10.4.1	乘	X	X					X
10.4.2	兼	X	X				X	X
10.7.1	門					X		X
11.1.1	雀							X
11.1.2	善				X			X
11.2.1	畢		X		X			X
11.3.1	麥	X				X[5]		X
11.4.1	爽	X						X
12.3.7	兩				X			X
12.8.1	堯		X		X	X[1]	X	X
13.1.1	黽		X		X			X
13.2.1	肅	X	X		X			X
13.7.1	鼠	X			X	X[5]		X

NUM	CHAR	W	E	N	S	B	I	F	NUM	CHAR	W	E	N	S	B	I	F
13.8.1	鼎				X			X									
14.4.1	爾	X	X	X			X	X									
14.4.2					X		X	X									
15.4.1			X					X									
16.1.1			X					X									

2.3 Variation

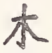
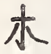
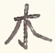
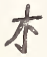
VARIATION PROCESS 1, (BORDER)

For certain components whose base forms have or in their "eastern" portions, this process causes the BORDER variant to assume a shape wherein becomes  and becomes  .

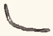
VARIATION PROCESS 2, (WEST)

For certain components which have in their "south central" portions, this process causes the WEST variant to assume a shape wherein becomes .

VARIATION PROCESS 3, (WEST)

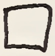
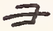
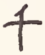
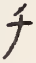
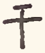
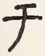
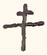
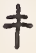
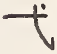
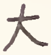
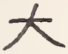
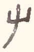
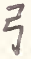
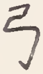
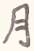
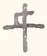
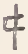
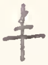
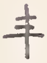
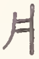
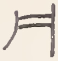
For certain components which contain  or a  -like structure, this process causes the WEST variant to assume a shape wherein  becomes  .

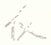
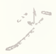
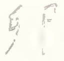
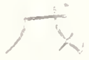
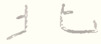
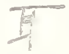
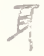
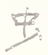
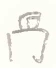
VARIATION PROCESS 4, (BORDER)

For certain components which have or in their "south-eastern" portions, this process cuases the BORDER variant to assume a shape wherein or become  .

Ad hoc Variations

Number	Base Form	Position	Variant
1.1.1	一	N (sometimes)	一
1.8.2	乙	B	乙
2.1.2	二	B	二
2.2.3	十	B	𠂇
2.4.2	乂	N	㐅
		I (sometimes)	㐅
		S (sometimes)	㐅
2.4.3	人	N	𠆢
2.4.4	八	N,S	ハ
2.5.3	入	I (sometimes)	入
2.5.11	丆	B	丆
2.5.13	𠂇	B	𠂇
2.7.1	丁	B	丁
2.8.14	乃	B	乃
3.1.4	子	W	子

Number	Base Form	Position	Variant F
3.1.6		B	
3.1.12		sometimes	
3.2.2		B,W (sometimes)	
3.2.3		B,S (sometimes)	
3.2.6		I (sometimes)	
3.2.7		S	
3.3.1		B	
3.4.1		N (sometimes)	
3.5.7		N (sometimes)	
3.8.7		B	
4.1.5		S	
4.1.18		sometimes	
4.2.8		E	
4.2.13		N,S	
4.2.15		B	
		I	

Number	Base Form	Position	Variant
4.3.6	戈	B	
4.4.9	氏	S	
4.4.14	朮	S (sometimes)	
4.5.8	豕	B	
4.5.10	少	B	
5.2.11	卵	B	
5.3.6	戊	B	
5.8.9	北	B	
6.2.6	耳	B	
		W	
6.3.10	虫	W	
6.7.2	四	B	
7.3.5	缶	S (sometimes)	

3. Concluding Remarks

This grammar is an attempt to characterize component combination from a particular point of view. The three modes of component combination which are utilized are precisely the productive ones. It is an experimental grammar and not a candidate for final success. The following remarks are intended to evaluate the grammar as such an attempt.

First, the grammar generates all the characters in Mathews' (2)[6] and every other unabbreviated Chinese character that we know of. Consequently, it constitutes a very compact defining device for a large complicated language. However, the grammar overgenerates its target language in the sense that some permitted sequences of rules result in the generation of structures which are clearly not acceptable Chinese characters. In future attempts, further constraints could be imposed along the lines discussed on page 5,ff in Rankin et. al. (4) to reduce this unacceptable output.

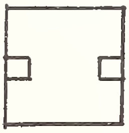
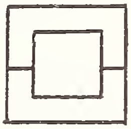
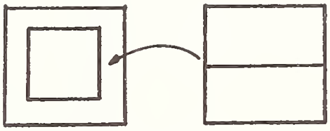
Second, the grammar preserves the distinction between component combination and stroke combination mentioned earlier. It does this by allowing only three modes of combination each of which operates on spatially disjoint arguments. It disallows such operations as touching and crossing of arguments which are characteristic of stroke combination.

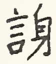
There is an accompanying disadvantage to disallowing touching and crossing because some instances of component combination cannot be explained by the grammar. For example, by disallowing crossing, we preserve the integrity of 十 , which is good, but also have to treat 東

---

[6] Barring clerical errors in the lexicon.

as a component, which is unfortunate, since 朩[7] and 日 are components themselves. There are quite a few "components" like 東 , which can be further segmented into smaller components.

Third, our process of frame-embedding produces a very simple, regular class of frames. There is a disadvantage that comes with this simplicity, however, and that is a certain lack of coverage. A fairly large number of multi-component structures is left unexplained due to the simplicity feature. Examples are 來 (木, 人, 人) which might require some such frame as  , and 是 (日 , 疋) as a border. Segmenting 是 would require embedding in the border subframe perhaps yielding  from  .

Fourth, the grammar generates some output characters in more than one way. For example, 謝 is generated as 言 plus 射 and also as  plus 寸 . We feel that most of these ambiguities correctly reflect the state-of-the-art of the linguistic analysis of Chinese characters in that we are not in all cases ready to select one analysis of a character as being correct and have the grammar suppress all other analyses. See Rankin et.al. (4) pages 56 ff for a detailed discussion of ambiguity problems.

Future grammars might examine the internal structure of components in the present lexicon in order to segment them if they contain substructures which are themselves already in the lexicon. This would require recognition of types of component combination which are far less common than those discussed in this paper. In particular, this would involve recognizing various types of superimposition and other component combination possibilities such as those underlying 來 and 是 discussed above. It is thought that one result of this work might be the development of complex frames which would correspond to instances of superimposition and other rare combination possibilities.

---

[7] 朩 is a variant form of 木 , component number 4.4.13.

We wish to thank Stephen J. Tauber, Justin A. Walker, and Russell A. Kirsch for reading earlier versions of this note and for making suggestions for improvement. Special thanks go to Stephanie Siegel for many valuable discussions of the material presented here and for many suggestions for improving the final version. Finally, we wish to thank Carrie L. Nuttall for typing the final version of this note.

4. References

1. Burkart, E.I., A Procedure for Decomposing Chinese-Japanese Ideographs, University of Pennsylvania Dissertation, 1968.

2. Mathews, R. H., A Chinese-English Dictionary, Cambridge, Massachusetts: Harvard University Press, 1960.

3. Rankin, B. K. III, A Linguistic Study of the Formation of Chinese Characters, University of Pennsylvania Dissertation, 1965.

4. Rankin, B. K. III, Siegel, S., McClelland, A., and Tan, J. L., A Grammar for Component Combination in Chinese Characters, Washington, D.C.: National Bureau of Standards Technical Note 296, 1966.

5. Rankin, B. K. III, Sillars, W. A., and Hsu, R. W., On the Pictorial Structure of Chinese Characters, Washington, D.C.: National Bureau of Standards Technical Note 254, 1965.

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH reports National Bureau of Standards research and development in physics, mathematics, chemistry, and engineering. Comprehensive scientific papers give complete details of the work, including laboratory data, experimental procedures, and theoretical and mathematical analyses. Illustrated with photographs, drawings, and charts.

Published in three sections, available separately:

• Physics and Chemistry

Papers of interest primarily to scientists working in these fields. This section covers a broad range of physical and chemical research, with major emphasis on standards of physical measurement, fundamental constants, and properties of matter. Issued six times a year. Annual subscription: Domestic, $9.50; foreign, $11.75*.

• Mathematical Sciences

Studies and compilations designed mainly for the mathematician and theoretical physicist. Topics in mathematical statistics, theory of experiment design, numerical analysis, theoretical physics and chemistry, logical design and programming of computers and computer systems. Short numerical tables. Issued quarterly. Annual subscription: Domestic, $5.00; foreign, $6.25*.

• Engineering and Instrumentation

Reporting results of interest chiefly to the engineer and the applied scientist. This section includes many of the new developments in instrumentation resulting from the Bureau's work in physical measurement, data processing, and development of test methods. It will also cover some of the work in acoustics, applied mechanics, building research, and cryogenic engineering. Issued quarterly. Annual subscription: Domestic, $5.00; foreign, $6.25*.

TECHNICAL NEWS BULLETIN

The best single source of information concerning the Bureau's research, developmental, cooperative and publication activities, this monthly publication is designed for the industry-oriented individual whose daily work involves intimate contact with science and technology—*for engineers, chemists, physicists, research managers, product-development managers, and company executives.* Annual subscription: Domestic, $3.00; foreign, $4.00*.

NONPERIODICALS

Applied Mathematics Series. Mathematical tables, manuals, and studies.

Building Science Series. Research results, test methods, and performance criteria of building materials, components, systems, and structures.

Handbooks. Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications. Proceedings of NBS conferences, bibliographies, annual reports, wall charts, pamphlets, etc.

Monographs. Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

National Standard Reference Data Series. NSRDS provides quantitive data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated.

Product Standards. Provide requirements for sizes, types, quality and methods for testing various industrial products. These standards are developed cooperatively with interested Government and industry groups and provide the basis for common understanding of product characteristics for both buyers and sellers. Their use is voluntary.

Technical Notes. This series consists of communications and reports (covering both other agency and NBS-sponsored work) of limited or transitory interest.

Federal Information Processing Standards Publications. This series is the official publication within the Federal Government for information on standards adopted and promulgated under the Public Law 89–306, and Bureau of the Budget Circular A–86 entitled, Standardization of Data Elements and Codes in Data Systems.

CLEARINGHOUSE

The Clearinghouse for Federal Scientific and Technical Information, operated by NBS, supplies unclassified information related to Government-generated science and technology in defense, space, atomic energy, and other national programs. For further information on Clearinghouse services, write:

Clearinghouse
U.S. Department of Commerce
Springfield, Virginia 22151

---

* Difference in price is due to extra cost of foreign mailing.

Order NBS publications from:

Superintendent of Documents
Government Printing Office
Washington, D.C. 20402

U.S. DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20230

OFFICIAL BUSINESS

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMMERCE